CHAPTER

Audio-Motor Integration for Robot Audition[1]

2

Antoine Deleforge[*], Alexander Schmidt[**] and Walter Kellermann[**]

[*] *Inria Nancy - Grand Est, 615 Rue du Jardin-Botanique, 54600 Villers-lès-Nancy, France*
[**] *Multimedia Communications and Signal Processing, Telecommunications Laboratory, University Erlangen-Nuremberg, Cauerstrasse 7, 91058 Erlangen, Germany*
[*] *Corresponding:* antoine.deleforge@inria.fr

ABSTRACT

In the context of robotics, audio signal processing *in the wild* amounts to dealing with sounds recorded by a system that moves and whose actuators produce noise. This creates additional challenges in sound source localization, signal enhancement and recognition. But the specificity of such platforms also brings interesting opportunities: can information about the robot actuators' states be meaningfully integrated in the audio processing pipeline to improve performance and efficiency? While robot audition grew to become an established field, methods that explicitly use motor-state information as a complementary modality to audio are scarcer. This chapter proposes a unified view of this endeavour, referred to as audio-motor integration. A literature review and two learning-based methods for audio-motor integration in robot audition are presented, with application to single-microphone sound source localization and ego-noise reduction on real data.

Keywords: audio-motor integration, robot audition, sound localization, ego-noise reduction, dictionary learning, single-channel, multi-channel.

[1] This is a preprint version of Chapter 2 - Audio-Motor Integration for Robot Audition in *Multimodal Behavior Analysis in the Wild*, Academic Press, 2019, Pages 27-51.

2.1 INTRODUCTION

The most natural way for humans to communicate is speech. For this reason, building robots which can interact with humans via speech is an important goal in robotics which has received growing research interest over the past 20 years [28, 43, 44, 56, 46, 36, 2, 57, 37]. Examples of desired high-level features for such communicating robots are the ability to look towards the person they interact with [58], speech recognition in noisy, multi-source and reverberant environments [69] or *speech diarization i.e.*, the identification of who talks to whom and when [51]. These high-level abilities can be associated to lower level audio signal processing tasks such as sound source localization and tracking [2, 57], sound source separation and speech enhancement [21] or dereverberation [47]. None of these tasks is specific to robot audition and they have been extensively studied over the past decades in contexts as varied as hearing aids, voice-controlled assistants in cars, smartphones or smart homes, audio signal restoration, or live music recording. But what makes robot audition fundamentally different from these fields?

A distinctive feature is that, by definition of a robot, the microphones are mounted on a system equipped with *actuators*, *i.e.*, motors. These actuators may impact received auditory signals in two different ways:

1. They may **change the microphone positions** if the latter are mounted on a mobile part such as a humanoid robot's head,
2. They may **create acoustic noise** at the microphones, referred to as *ego-noise* .

Sensor mobility may be used as an asset by actively placing sensors in order to improve sound source localization [43, 55, 9, 48] or speech enhancement [45, 4, 67]. This is referred to as *active audition*. On the other hand, moving sensors also prevent the use of classical audio signal processing tools that assume a static sound propagation model from sources to microphones, such as beamforming [21]. Besides, ego-noise may significantly impair auditory scene analysis, especially when microphones are placed near the actuators [36, 75]. But contrary to other audio signal processing applications involving noise and movement, robots can benefit from *proprioceptors*. These sensors provide information on the current *motor-state*, *i.e.*, the translational and rotational position, speed and acceleration of the robot's actuators. Can this additional modality be used to the benefit of robot audition in unconstrained environment? Despite its potential usefulness, methods explicitly integrating it to the audio processing pipeline are scarce [33, 20, 60, 61]. We refer to this endeavour as *audio-motor integration*.

This chapters propose a unified view on audio-motor integration methodologies for robot audition. Section 2.2 starts by a literature overview of related works in the fields of psychophysics and robotics. Then, two learning-based audio-motor integration models and their applications to real-world tasks are presented. Section 2.3

introduces a single-channel sound source localization method based on head movements. Section 2.4 presents a general ego-noise reduction framework exploiting proprioceptors and dictionary learning. Finally, a conclusion and future perspectives on audio-motor integration for robot audition is presented in Section 2.5. An illustration of the different components of a typical auditory-motor system, their associated literature, and the parts addressed in this chapter is showed in Fig. 2.1.

2.2 AUDIO-MOTOR INTEGRATION IN PSYCHOPHYSICS AND ROBOTICS

The majority of studies on audio-motor integration, both in psychophysics and robotics, focus on sound source localization. From a psychophysical point of view, it is known since the experiments of Lord Rayleigh in 1907 that humans use *binaural* cues in order to estimate the direction of a sound source [64]. Two types of binaural cues seem to play an essential role, namely *interaural level differences* (ILD) and *interaural time differences* (ITD), the frequency domain counterpart of the latter being referred to as *interaural phase differences* (IPD). Both ILD and IPD are known to be subject-dependent and frequency-dependent cues. This is captured by the so-called head related transfer functions (HRTFs) determined by the shape

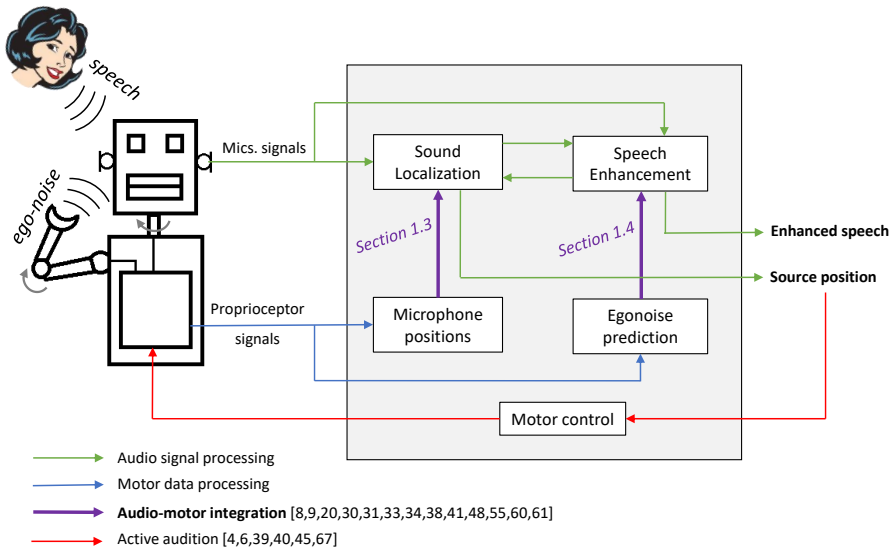

FIGURE 2.1 Auditory-Motor System Components

Illustration of an auditory-motor system whose goal is to enhance a speech source and to turn the head towards it in the presence of ego-noise.

of the head, pinna and torso. HRTFs filter the sound propagating from a source to eardrums and depend on the source direction. It is known that spatial information provided by interaural-difference cues within a restricted band of frequency is spatially ambiguous, particularly along vertical and front/back axis [42]. This suggests that humans make use of full spectral information for 2D sound source localization [25]. This is confirmed by biological models of the auditory system hypothesizing the existence of neurons dedicated to the computation of interaural cues in specific frequency bands [74]. A large number of computational models were developed for robust sound localization and tracking based on ITD, ILD, IPD and HRTFs in the context of robot audition, see for example the recent reviews [2, 57].

While the features mentioned above are generally used assuming static sensors, sound localization features could also be extracted using sensor motions, although this idea has received much less attention. Early psychophysical experiments suggested that head motions are useful for disambiguating potential confusions generated by the human pinna's filter [73], notably to estimate the elevation of low frequency sounds [53]. Other experiments [66, 76] further support the idea that head movements are useful for localization, although less significantly so than ITD and ILD [42]. Interestingly, [32] suggests that taking into account head motions is useful to improve the impression of *3D sound* in virtual reality. They show that when listening through different HRTFs than his/her own, a listener often complains that auditory events are spatially diffuse, and makes incorrect judgements about the source locations. The experimental study of [32] on human subjects demonstrates that head motion can overcome HRTF mismatches, increasing perceived location accuracy.

Despite these psychophysical evidences, very few computational sound localization studies incorporate head motion. In [33], several ITD values obtained from different motor states with a two-microphone device with two degrees of freedom (translation, rotation) are used for 2D sound source direction and distance estimation. The authors of [41, 38] propose to map binaural cues to the azimuths of multiple sound sources using Gaussian mixture models [41] or deep neural networks [38] and two static head positions are used to resolve front-back ambiguities. In [55, 9, 48] mobile robots are used to actively collect several viewpoints of one or several emitting sound sources in order to accurately estimate their azimuths and distances. More broadly, studies tackling the problem of audio-based simultaneous localization and mapping (SLAM) have recently emerged [19, 34]. Interestingly, [34] does it with a single microphone at multiple viewpoints thanks to acoustic echoes. Note that the above mentioned studies rely on static localization techniques applied from different viewpoints. In contrast, [8] computes the average binaural cross-correlation of a continuously rotating head to estimate a source's azimuth in an anechoic scenario. To the best of the authors' knowledge, exploiting continuous sensor movement for 2D sound sound direction estimation with a single microphone, as showed in Section 2.3 of this chapter, has not been proposed before.

Complementarily to sound source localization, [45] introduced the idea of *active*

robot audition (see Fig. 2.1) where the head of a robot is directed towards estimated source locations in order to enhance the signal of interest via, *e.g.* beamforming. This idea was recently further exploited in [4, 67] using a so-called *robomorphic array*, where microphones are placed on a robot's limbs. In [39], it is showed that turning the head of a binaural robotic system towards an estimated source direction enhances sound localization performance. From a psychophysical viewpoint, it is well known that humans use active audition to improve sound perception notably to address the so-called *cocktail party problem* [24]. Other approaches focus on the goal of orienting a system towards a sound source, without necessarily exploiting this for enhancement. This ability is referred to as phonotaxis, and was implemented on a rat-inspired robot equipped with mobile ears in [6]. More recently, sound has been used in feedback-control loops, *i.e.*, *audio-servoing*. In [40] motor commands are used to align perceived ITDs to a target value, resulting in the robot placing itself on a line with constant ITD value.

Another category of methods consists in using the motor capabilities of a robot to help it *learn* sound source localization techniques [5, 27, 65, 15, 7, 12], much like humans and mammals do during early stages of development. Several psychophysical studies suggest that the link between auditory features and source locations would not be hard-coded in the brain but rather learned [77, 3] or re-learned [26] from experience. One example of such learning processes is the sensori-motor theory of perception, originally laid by Poincaré [54] and more recently investigated in [49]. This theory suggests that experiencing the sensory consequences of voluntary motor actions is necessary for an organism to learn the perception of space. In [3] a psychophysical sensorimotor model of sound source localization using head-related transfer function (HRTF) datasets of bats and humans was proposed. Similar ideas were implemented on robots using reinforcement learning [5], linear regression [27], locally-linear regression [12], look-up tables [15] or manifold learning [65, 7]. Interestingly, [65] uses multiple view points and diffusion kernels to learn the estimation of the azimuth angle of a white-noise source using a single microphone.

Finally, a few robotic studies use audio-motor integration for *ego-noise* reduction. The idea is to map the current motor-state of a robot to some estimates of the corresponding acoustic noise produced by the motors. For instance, the ego-noise power-spectral density can be predicted this way using an artificial neural network [31] or a nearest neighbour search [30] based on a training dataset. In [20] the same idea is applied to predict multichannel ego-noise covariance matrices using Gaussian process regression. These statistics can be used to cancel the noise via Wiener filtering, under a local stationarity assumption. In [60, 61], a motor data-guided dictionary-learning framework is presented, allowing to model noise which is non-stationary both spatially and spectrally. This framework and exemplary results on real data are presented in Section 2.4 of this chapter.

2.3 SINGLE-MICROPHONE SOUND LOCALIZATION USING HEAD MOVEMENTS

In this section, we present a method that uses motor movements to extract auditory features enabling 2D sound source localization with a single microphone placed on an acoustic dummy head[1]. The mapping from features to direction is learned from a training dataset, building on the audio-motor, learning-based sound source localization framework of [12].

2.3.1 HRTF MODEL AND DYNAMIC CUES

Let $s(f,t) \in \mathbb{C}$ denote the signal emitted by a static sound source and $y(f,t) \in \mathbb{C}$ denote the signal recorded by a microphone placed on an acoustic dummy head in the short-time Fourier domain, with f and t denoting frequency and time indexes, respectively. We assume that the emitted signal is stationary over the considered time interval. The acoustic head acts as a rigid body causing reflections and shadowing of the source signal. The corresponding linear filter, referred to as the head related transfer function (HRTF), is denoted $a(f,\phi,\psi)$ and depends on the frequency f and on the relative azimuth ϕ and elevation ψ of the source in the microphones' frame[2]. Note that in the case of microphone movements, ϕ and ψ will depend on time as well. We have:

$$y(f,t) = a(f,\phi(t),\psi(t))s(f,t). \tag{2.1}$$

We also define the *received log-power* spectrogram by

$$p(f,t) = \log|y(f,t)|^2 = \log|a(f,\phi(t),\psi(t))|^2 + \log|s(f,t)|^2. \tag{2.2}$$

Since the emitted signal $s(f,t)$ is assumed stationary, its power spectral density $\mathbb{E}_t\{|s(f,t)|^2\}$ does not depend on time, and hence its instantaneous estimate $|s(f,t)|^2$ should be independent of t on average.

Let us now consider that the microphone is mounted on a robotic head with 2 degrees of freedom: pan (azimuth rotation) and tilt (elevation rotation), such as the one showed in Fig. 2.2. Any motor command $\boldsymbol{\xi}$ will change the relative direction (ϕ,ψ) of the source in the microphone's frame. A command $\boldsymbol{\xi}$ can be identified with a function associating a time t to a motor state $\boldsymbol{\xi}(t)$. This motor state can be accessed through the proprioceptors of the robot. We define a *dynamic cue* $\boldsymbol{\tau}(\boldsymbol{\xi}) = \{\tau(f,\boldsymbol{\xi})\}_{f=1}^{F} \in \mathbb{R}^F$ by the *expected temporal derivatives* of $p(f,t)$ at all frequencies

[1] This study extends the unpublished technical report [14] by the first author.

[2] Assuming that the sound source is placed in the far field, the dependency on source distance can be neglected. As showed in [50], this typically occurs for distances $> 1.8m$ on binaural systems.

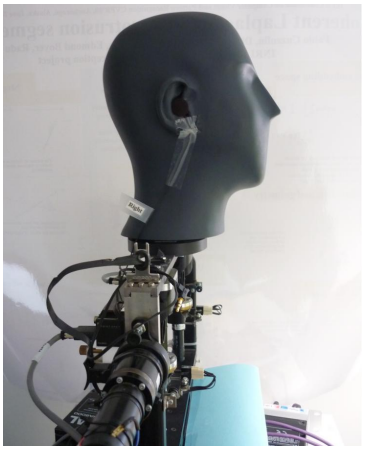

FIGURE 2.2 Active Binaural System

The POPEYE setup used in the CAMIL dataset [14] and in our experiments. A binaural dummy head is mounted on a motor system with two degrees of freedom: pan (left-right) and tilt (up-down).

while the robot performs command $\boldsymbol{\xi}$:

$$\tau(f,\boldsymbol{\xi}) = \mathbb{E}_t\left\{\frac{\partial p(f,t)}{\partial t}\right\},\ f = 1\ldots F, \tag{2.3}$$

where F is the number of frequency bands considered. As mentioned, the stationarity assumption on $s(f,t)$ implies that $\log|s(f,t)|^2$ is time-independent on average and hence its expected derivative is zero. On the other hand, the motor command $\boldsymbol{\xi}$ significantly changes the relative source directions ϕ and ψ, implying that $\log|a(f,\phi(t),\psi(t))|^2$ has a non-negligible expected derivative. Hence, $\tau(f,\boldsymbol{\xi})$ is a spatial cue that does not depend on the emitted signal and only depends on the source direction (ϕ,ψ).

In practice, $\boldsymbol{\tau}(\boldsymbol{\xi})$ can be approximated at each frequency by the slope of the least-square linear regression between discrete time indexes and the received log-energies $\{p(f,t)\}_{t=1}^{T}$, where T is the number of time frames considered. The validity of this approximation relies on the fact that typical HRTFs are approximately locally-linear with respect to source directions, as demonstrated in [12]. Hence, using motor commands with constant angular velocities, $p(f,t)$ should vary approximately linearly in time, justifying the use of linear regression. The slope of the least-square linear regression is then a least mean square estimator of the expected temporal derivative, assuming zero-mean perturbations. This is illustrated in Figure 2.3 using the CAMIL dataset version 0.1[3] [14]. This dataset was recorded with the setup of Fig.

[3] Data available at`https://team.inria.fr/perception/the-camil-dataset/`.

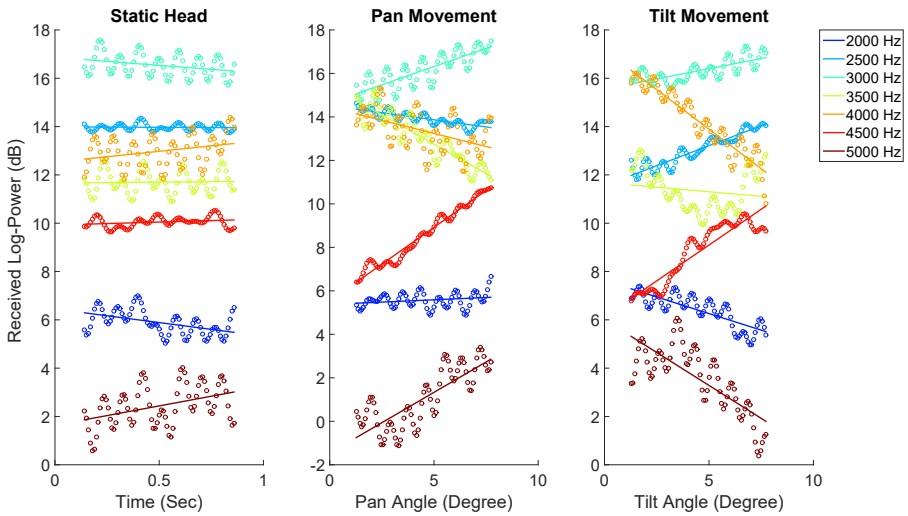

FIGURE 2.3 Illustration of Dynamic Cues

These figures represent the received log-power (2.2) by a single acoustic dummy-head microphone as a function of time (or head-angle) at different frequencies. The least-square linear regression of each curve is represented by a solid line. The emitted signal is an 800ms random mixture of 600 sine waves with frequencies ranging from 50Hz to 6000Hz. Left: the head is static. Middle: the head performs a pan movement at constant $9°/sec$ velocity. Right: the head performs a tilt movement at constant $9°/sec$ velocity.

2.2. As can be seen, the stationarity of the emitted signal implies that the received log-power is roughly constant over time when the head is static (left). On the other hand, head movements induce near linear variations of the log-power. The expected derivative of these variations can be approximated by the slope of their least-square linear regression (solid lines). The sampling frequency of the signal was 48 kHz and the sliding short-time Fourier Hamming window was set to 200 ms with 95% overlap. This resulted in 8193 discrete positive frequencies between 0 and 24 kHz and $T = 101$ time frames per second of signal. In practice, only cues corresponding to frequencies between 1500 and 6000 Hz were kept, as they showed to be the most useful for localization. This resulted in $F = 1537$ frequency bands used in the proposed dynamic cues.

2.3.2 LEARNING-BASED SOUND LOCALIZATION

Once dynamic cues $\boldsymbol{\tau}(\boldsymbol{\xi})$ are computed, they need to be mapped to a corresponding source direction. While computationally demanding physical models exist to generate HRTF filters based on source directions and accurate 3D head models [78], we are interested here in the other way around mapping, *i.e.*, from dynamic cues to source

directions. This cannot be easily obtained in practice, in particular when head movements are considered. Due to the infeasibility of fully modelling the physics of sound propagation in realistic settings, an alternative approach has recently emerged and is referred to as *supervised* or *learning-based* sound source localization [65, 12, 16]. These methods bypass the use of an explicit, approximate physical model by directly learning a mapping from audio features to spatial properties using an appropriate training dataset.

The CAMIL dataset version 0.1 [14] consists of recordings made with a binaural dummy head (Fig. 2.2) in the presence of a sound source (loudspeaker) placed at 16, 200 annotated relative directions in the microphones' frame. The source is placed 2.7 meters from the receiver in all recordings, and the reverberation time of the room is around 400ms. For each direction, 3 recordings of 1 second each are available: (i) no head movement, (ii) a 9° pan movement rightwards at constant speed and (iii) a 9° tilt movement downwards at constant speed. The emitted source signals are designed to be approximately stationary and correspond to random linear combinations of 600 sine waves with frequencies ranging from 50Hz to 6000Hz and random phase offsets. During motor commands, the annotated relative source position corresponds to the one half-way through the movement. For the single microphone experiments, only the left microphone channel is used.

The training dataset is composed of N pairs $\{\boldsymbol{\tau}_n(\boldsymbol{\xi}), \boldsymbol{z}_n\}_{n=1}^N \subseteq \mathbb{R}^F \times \mathbb{R}^2$ where $\boldsymbol{z}_n = (\phi_n, \psi_n)$ is the $n-$th source direction and $\boldsymbol{\xi}$ is a fixed command throughout the dataset. Three motor commands are considered: *pan*, *tilt* or the successive combination of both, each at constant angular velocity $(9°/sec)$. Only source directions corresponding to azimuth angles between $-90°$ and $+90°$ and elevation angles between $-45°$ and $+45°$ are kept, resulting in 3812 directions out of which $N = 2859$ were kept for training and the 953 others for testing.

This training set must be used so that given a new test observation $\tilde{\boldsymbol{\tau}}(\boldsymbol{\xi})$, an associated source direction $\tilde{\boldsymbol{z}}$ can be estimated. To achieve this, we use the high- to low-dimensional regression method Gaussian locally-linear mapping (GLLiM[4] [17]) proposed in [13]. GLLiM is a probabilistic method that estimates Q local affine transformations from a low-dimensional space (here, the space of source directions) to a high-dimensional space (here, the space of dynamic cues) using a Gaussian mixture model. This mapping is then reversed through Bayesian inversion, yielding an efficient estimator of $\tilde{\boldsymbol{z}}$ given $\tilde{\boldsymbol{\tau}}(\boldsymbol{\xi})$. GLLiM was notably successfully applied to supervised binaural sound source localization using either real [16, 12] or simulated [22] training sets. Here, a fixed value $Q = 50$, diagonal and equal noise covariance matrices and equal mixture weights are used in all experiments (see [13] for details on the GLLiM method).

[4] The code for this method is available at `https://team.inria.fr/perception/gllim_toolbox/`

2.3.3 RESULTS

The proposed method was trained on a random subset of $N = 2859$ cue-to-direction pairs and tested on the remaining 953 dynamic cues. This was done so that the emitted training sounds are all distinct from the emitted test sounds and the training source directions are distinct from the test source directions. Results are shown in Table 2.1. As can be seen, audio-motor integration and small head movements enable the localization of a sound source in 2D with high precision ($< 4°$ with combined commands) using a *single microphone*. This is impossible to achieve with any of the existing sound source localization methods in the literature, to the best of the authors' knowledge. For comparison, results obtained with a static head and dynamic cues (1 microphone) or binaural cues (ILD and IPD, 2 microphones, as in [12]) are showed in Table 2.2. Unsurprisingly, using dynamic cues with a static head yields localization results similar to randomness, with over 50% of outliers. This is because the absence of movement removes spatial information from $\boldsymbol{\tau}(\boldsymbol{\xi})$, bringing (2.3) close to 0. On the other hand, using traditional binaural cues with two static microphones, as done in *e.g.*, [12], yields comparable results to the ones obtained with the proposed dynamic cues and a single moving microphone. This validates the feasibility of single-microphone sound source localization by audio-motor integration.

An important limit of this approach is that it requires a carefully annotated training dataset, which is likely to be room- and system-dependent. Besides, the method strongly relies on the assumption that emitted signals are approximately stationary during the emission period ($\approx 0.8s$ in our experiments). Robust extensions of this method that discard time-frequency points with low-energy would be an interesting route to investigate. Besides engineering applicability, these results are the first ones to corroborate psycho-physical evidence suggesting that small continuous head movements may help localizing sounds [73, 66, 53, 76], using an artificial system.

TABLE 2.1 Single-microphone 2D dynamic sound localization results

Azimuth and elevation estimation errors on the testing set using a single microphone and different motor commands. Results are presented in the form Avg±Std (Out %) where Avg±Std denote the mean and standard deviations of *inlying* absolute angular errors and Out denotes the percentage of *outliers*. Outlying errors are defined as those larger than 30°.

Command:	Pan	Tilt	Pan + Tilt
Azimuth (°)	3.00 ± 3.3 (0.5%)	4.02 ± 3.5 (0.1%)	3.26 ± 3.2 (0.0%)
Elevation (°)	2.50 ± 2.8 (0.3%)	1.62 ± 1.3 (0.2%)	1.71 ± 1.6 (0.0%)

TABLE 2.2 2D static sound localization results

Azimuth and elevation estimation errors on the testing set using one static microphone and dynamic cues or two static microphones and binaural (ILPD) cues. Results are presented in the same form as Table 2.1.

Method:	Dynamic cues (1 mic.)	ILD + IPD [12] (2 mic.)
Azimuth (°)	14.0 ± 8.7 (65%)	2.04 ± 1.6 (0.0%)
Elevation (°)	13.5 ± 8.5 (40%)	1.26 ± 1.0 (0.0%)

2.4 EGO-NOISE REDUCTION USING PROPRIOCEPTORS

This section presents a general audio-motor integration framework for ego-noise reduction, initially presented in [60] and extended in [61]. This frameworks enables the modelling of noise signals that are non-stationary both spatially and spectrally by building on a phase-optimized dictionary-learning method[5] [17] which is also outlined here for completeness.

2.4.1 EGO-NOISE: CHALLENGES AND OPPORTUNITIES

When a robot is moving, its rotating joints as well as the moving parts of its body cause significant noise which is referred to as ego-noise. This corrupts recordings and therefore degrades performance of, e.g., a speech recognizer. To relieve this problem, a suitable noise reduction mechanism is required. This task is particularly challenging because the noise involved is often louder than the signals of interest. Moreover, it is highly non-stationary as the robot performs different movements with varying speeds and accelerations. Furthermore, ego-noise cannot be modelled as a single static point interferer as the joints are located all over the body of the robot.

All this discourages the use of traditional statistical noise reduction techniques such as (multichannel) Wiener filtering or beamforming [36, 21]. On the bright side, however, two robot-specific opportunities may be exploited. First, ego-noise will usually be strongly structured both spatially and spectrally (*e.g.*, Fig. 2.4, bottom-right) because it is produced by an automated system restricted to a limited number of degrees of freedom. Second, important extra information may be exploited in addition to the audio signals, namely, the instantaneous motor state of the robot, e.g., the joints' angles and angular velocities collected by proprioceptors (*e.g.*, Fig. 2.4, top-right).

[5] The code for this method is available at `https://robot-ears.eu/po_ksvd/`

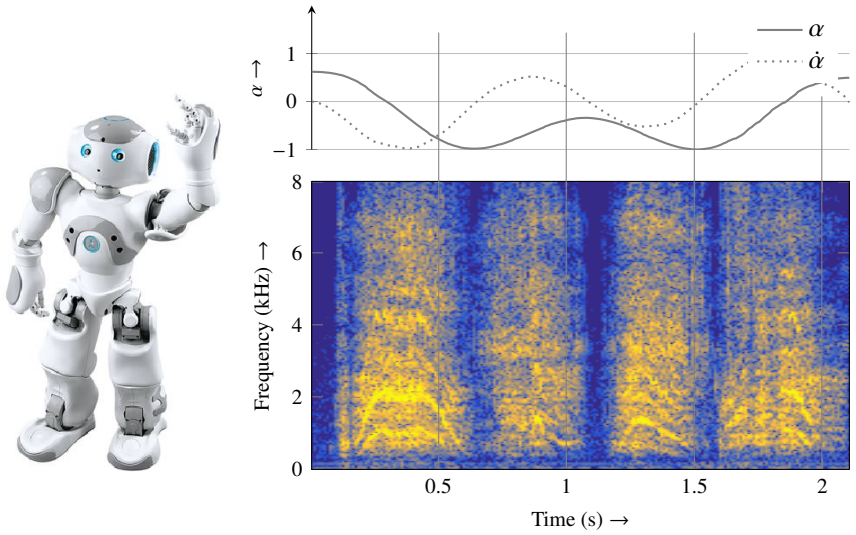

FIGURE 2.4 Ego-noise of the humanoid robot NAO waving the arm

Left: the robot NAO. Bottom-right: Ego-noise Spectrogram of Nao waving the left arm, showing distinctive spectral structures. Note that stationary fan noise components have been removed by a multi-channel Wiener filter [36]. Top-right: corresponding motor data, i.e., the angle α and first derivative $\dot{\alpha}$ of the left shoulder pitch joint involved in the movement (both normalized).

2.4.2 PROPRIOCEPTOR-GUIDED DICTIONARY LEARNING

The existence of a strong spectral structure in ego-noise motivates the use of dictionary learning methods. The idea is to express structure in terms of sparsity in a particular basis. More precisely, if $\mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T] \in \mathbb{C}^{P \times T}$ represents T examples of P-dimensional signals, there must exist a set of K atoms or a *dictionary* $\mathbf{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K] \in \mathbb{C}^{P \times K}$ such that each signal is a linear combination of only a few atoms, *i.e.*, $\mathbf{Y} \approx \mathbf{DX}$ where $\mathbf{X}$ has sparse columns. Estimating $\mathbf{D}$ and $\mathbf{X}$ from $\mathbf{Y}$ is a sparse instance of matrix factorization. In audio signal processing, it is natural to seek such a factorization in the non-negative power spectral density (PSD) domain, since the magnitude spectra of natural sounds such as speech often feature redundancy and sparsity. This approach gave rise to a large number of methods for audio signal representation and extraction within the framework of *non-negative matrix factorization* (NMF) [63, 72]. Extensions of NMF to complex-valued multichannel spectrograms have later been proposed [52, 59]. Most of these extensions assume a simple spatial structure: the modelled signal consists in a mixture of a few fixed, point sources, *i.e.*, with constant steering vectors. Such models are not appropriate for ego-noise which features complex spatial structures. In [17], a model including both complex spectral and spatial structures was proposed, via *phase-optimized*

dictionary learning. This model is summarized in Section 2.4.3 of this chapter.

The general concept of dictionary-based noise reduction comprises the following steps. First, a noise dictionary $\mathbf{D}_{\text{noise}}$ is learned from a set of noise-only examples $\mathbf{Y}_{\text{noise}}$ via the factorization $\mathbf{Y}_{\text{noise}} \approx \mathbf{D}_{\text{noise}}\mathbf{X}_{\text{noise}}$ with $\mathbf{X}_{\text{noise}}$ being maximally sparse. Then, given a test observation $\tilde{\boldsymbol{y}} \in \mathbb{C}^P$ containing both noise and target signals, its noisy part is estimated by finding $\tilde{\boldsymbol{x}} \in \mathbb{C}^K$ such that $\tilde{\boldsymbol{y}} \approx \mathbf{D}_{\text{noise}}\tilde{\boldsymbol{x}}$. The enhanced target signal is then given by $\tilde{\boldsymbol{y}} - \mathbf{D}_{\text{noise}}\tilde{\boldsymbol{x}}$.

Besides methods purely based on the structure of audio signals, another approach stipulates to exploit available proprioceptor, *i.e.*, motor data, and map them to a noise model. In [31], the time-varying noise power spectral density (PSD) is estimated by a deep neural network (DNN). The DNN is fed by motor data, which incorporates not only current, but also past sensor values. In [30], PSD noise templates are used for spectral subtraction. In the learning step, only ego-noise is present and each point in the motor data space is associated with a certain spectral noise template. In the testing (or working) phase, the approach uses a nearest-neighbour criterion to find the best matching template, which is then subtracted from the magnitude spectrum of the recording. In [20], multichannel noise covariance matrices are predicted from motor states via Gaussian process regression.

In [60], we introduced an alternative and generic audio-motor integration framework to fuse the information brought by an advanced learned structured audio model on the one hand, and instantaneous motor data on the other hand. The key idea is to replace the computationally costly search in the dictionary, which is untractable for large dictionaries (NP-hard [11]), by a classification procedure guided by current proprioceptor data $\boldsymbol{\alpha}$. These data are fed into support vector machines (SVMs) [62] to efficiently find suitable entries in the ego-noise dictionary $\mathbf{D}_{\text{noise}}$. We showed that this approach reduces computational complexity while simultaneously improving performance. This approach and some results are outlined in Section 2.4.4.

2.4.3 PHASE-OPTIMIZED DICTIONARY LEARNING

Although ego-noise is highly non-stationary, it has distinctive spectral and spatial characteristics. The basic idea of a dictionary representation is to capture such characteristics by a collection of prototype signals, called atoms, collected in a dictionary. In our case, the structured ego-noise signal should be represented by a linear combination of a few atoms at each time frame. If these atoms are specifically designed to represent signals sharing spectral and spatial characteristic of ego-noise only, subtracting these atoms should remove the noise while preserving the residual signal of interest such as speech. We briefly summarize here the recent approach [17] that automatically learns a multichannel dictionary capturing both spatial and spectral characteristics of a training signal. In the following we represent a multichannel signal in the spectral domain by concatenation of the M channels per frequency bin, giving a signal vector of dimension $P = MF$, where F represents the number of frequency

bins per channel. Then, we denote the dictionary by $\mathbf{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K] \in \mathbb{C}^{P \times K}$ containing K atoms $\boldsymbol{d}_k \in \mathbb{C}^P$. Moreover, the dictionary is corrected by a time-varying phase matrix $\boldsymbol{\Phi}_t \in \mathbb{C}^{F \times T}$ at each time frame t, where each element has unit complex modulus. The phase-corrected dictionary is then given by

$$\mathbf{D}\{\boldsymbol{\Phi}_t\} := \begin{pmatrix} \boldsymbol{d}_{1,1} & \ldots & \boldsymbol{d}_{1,K} \\ \boldsymbol{d}_{2,1} & \ldots & \ldots \\ \vdots & \ddots & \vdots \\ \boldsymbol{d}_{F,1} & \ldots & \boldsymbol{d}_{F,K} \end{pmatrix} \odot \begin{pmatrix} \phi_{1,1,t} & \ldots & \phi_{1,K,t} \\ \phi_{2,1,t} & \ldots & \ldots \\ \vdots & \ddots & \vdots \\ \phi_{F,1,t} & \ldots & \phi_{F,K,t} \end{pmatrix} \tag{2.4}$$

where each element $\boldsymbol{d}_{f,k} \in \mathbb{C}^M$ captures the spectral value of atom k at frequency bin f as well as the relative phases and gains between the M channels. Here, $\odot$ denotes a modified Hadamard product, where each vector $\boldsymbol{d}_{f,k} \in \mathbb{C}^M$ in matrix $\mathbf{D}$ is multiplied by a global phase term $\phi_{f,k,t} \in \mathbb{C}$ in $\boldsymbol{\Phi}_t$. A given multichannel spectrogram frame $\boldsymbol{y}_t$ should then be approximated by $\boldsymbol{y}_t \approx \mathbf{D}\{\boldsymbol{\Phi}_t\}\boldsymbol{x}_t$, where the vector $\boldsymbol{x}_t \in \mathbb{C}^K$ picks the atoms from the dictionary. Since only a few atoms should be used, $\boldsymbol{x}_t$ is constrained to be sparse, *i.e.*, it should contain at most $S_{\max}$ nonzero elements, where $S_{\max}$ is referred to as the *sparsity level*. The overall problem can then be written as

$$\begin{aligned} &\underset{\mathbf{D},\boldsymbol{\Phi},\mathbf{X}}{\operatorname{argmin}} \sum_{t=1}^{T} \|\boldsymbol{y}_t - \mathbf{D}\{\boldsymbol{\Phi}_t\}\boldsymbol{x}_t\|_2^2 \quad \text{subject to:} \\ &\|\boldsymbol{x}_t\|_0 \leq S_{\max},\ x_{kt} \geq 0 \text{ and } |\phi_{ft,k}|^2 = 1\ \forall\ f, k, t. \end{aligned} \tag{2.5}$$

Here, $\|\cdot\|_2$ and $\|\cdot\|_0$ denote the ℓ_2- and ℓ_0-norm, respectively. The latter counts the number of nonzero elements in $\boldsymbol{x}_t$. The minimization (2.5) is done with respect to (w.r.t.) different arguments, depending on which stage of the algorithm is considered. The training and testing stages are outlined in the following (see [17] for details).

- **Training:** (2.5) is minimized w.r.t. $\mathbf{D}$, $\boldsymbol{x}_t$ and $\boldsymbol{\Phi}_t$. In the training stage, $\mathbf{D}_{\text{noise}}$ should be learned using a set of training examples $\mathbf{Y}_{\text{noise}}$. For this, [17] proposes the phase-optimized K-SVD (PO-KSVD) algorithm. It can be viewed as a phase-optimized complex extension of the popular dictionary learning method K-SVD [1]. It alternates between a sparse coding step and a dictionary update step.
- **Testing:** (2.5) is minimized w.r.t. $\boldsymbol{x}_t$ and $\boldsymbol{\Phi}_t$ for each new test observation $\tilde{\boldsymbol{y}}_t$ while the pre-trained dictionary $\mathbf{D}_{\text{noise}}$ is fixed. The best fitting entries from dictionary $\mathbf{D}_{\text{noise}}$ are searched and subtracted from $\tilde{\boldsymbol{y}}_t$, which may contain ego-noise and speech, for example. The NP-hard problem of finding the best combination of atoms is done using an extension of orthogonal matching pursuit (OMP, after [68]) due to its empirically good performance. The extension is called PO-OMP (for phase-optimized OMP).

2.4.4 AUDIO-MOTOR INTEGRATION VIA SUPPORT VECTOR MACHINES

2.4.4.1 Method description

While the knowledge of the robot's instantaneous motor state should intuitively be beneficial for ego-noise reduction, a crucial question is at which stage motor data should be included. As described in the previous section, one of the main bottlenecks of the multichannel dictionary method in [17] is the testing phase, where an NP-hard sparse coding problem is approximately solved by the costly iterative PO-OMP procedure. Hence, we propose to replace the entire testing stage by a novel and more efficient motor-guided atom selection method while keeping the dictionary learning stage of [17], for which computational time is less of an issue.

The physical state of a 1-dimensional robot joint can be described by its position in terms of an angle α_t at a given time stamp τ_t. Furthermore, from successive angle stamps we can calculate a discrete-time approximation of its first angle derivatives, i.e., angle speed

$$\dot{\alpha}_t = \frac{\alpha_t - \alpha_{t-1}}{\tau_t - \tau_{t-1}}.$$

For each joint, we collect the recorded and calculated angle data in a feature vector $\boldsymbol{\alpha}_t = (\alpha_t, \dot{\alpha}_t)$, which is, a bit loosely, referred to as motor data in the following. Note that using the joints' acceleration is also an option, but is disregarded here as it did not prove useful in our experiments. Assume a spectrum $\boldsymbol{y}_t$ is observed at a time frame τ_t in which ego-noise alone is present. Additionally, the motor data $\boldsymbol{\alpha}_t$ for this frame is available. The proposed concept stipulates to associate this motor data point with those atoms which PO-OMP would select in a pre-trained ego-noise dictionary to represent spectrum $\boldsymbol{y}_t$. Note that the order in which atoms are chosen is unimportant. For example, if at time step t PO-OMP selects atoms 8, 5 and 7 in the dictionary and at time step $t+1$ the output is 7, 8 and 5, both motor data samples $\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}_{t+1}$ are associated with the set of atom $\{8, 7, 5\}$. The curly brackets $\{\cdot\}$ emphasize that the order of selected atoms is not considered. The number of possible atom combinations for any given atom set is then

$$N_C = \binom{K}{S_{\max}}. \tag{2.6}$$

In the following, the N_C possible atom sets are denoted as $\hat{\mathbf{D}}_q \in \mathbb{C}^{P \times S_{\max}}, q = 1 \cdots N_C$, where the choice of notation emphasizes that each $\hat{\mathbf{D}}_q$ contains a selection of atoms from $\mathbf{D}$. Our plan in the following is to decide on a set $\hat{\mathbf{D}}_q$ based on a motor data sample $\boldsymbol{\alpha}_t$ and use all atoms in $\hat{\mathbf{D}}_q$ for ego-noise suppression.

Fig. 2.5 shows some motor data points in the $(\alpha, \dot{\alpha})-$plane which are associated to a certain set of atoms as an example. They appear to form clusters. The non-linearity

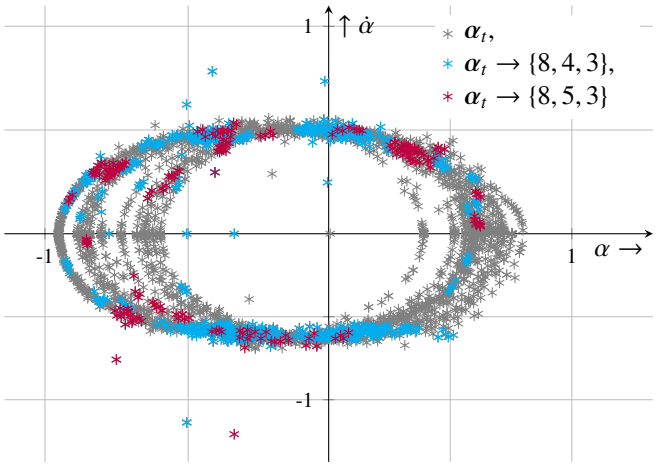

FIGURE 2.5 Clustering of used auditory atoms in the motor-state space

Motor data points $\boldsymbol{\alpha}_t = (\alpha_t, \dot{\alpha}_t)$ for a right shoulder pitch movement. Points highlighted in blue and red are those associated with a certain set of selected atoms in the dictionary. They appear to form clusters.

of the clusters' contours motivates the use of a kernel method. It is reasonable to classify these points using a classifier $C_q(\boldsymbol{\alpha}) \in \{-1, +1\}$ deciding if a new incoming motor data point $\boldsymbol{\alpha}$ falls into the clustering area of atom set q. If yes, $C_q(\boldsymbol{\alpha}) = 1$, if not $C_q(\boldsymbol{\alpha}) = -1$ holds. All in all, N_C such classifiers must be trained. For notational convenience, they are represented in vector form

$$C(\boldsymbol{\alpha}) = [C_1(\boldsymbol{\alpha}), \ldots, C_{N_C}(\boldsymbol{\alpha})]. \tag{2.7}$$

We propose to model the data points to cluster as following an unknown probability density function (pdf). By estimating its support, the above described clustering problem is solved. To do so, we use a method from the broad range of support vector machines (SVM). The 1-Class-SVM [62] is a method based on Gaussian kernels that estimates a classifier $C(\cdot)$ whose decision boundaries can be shown to be the support of a pdf that generated the training data with high probability. The reader is referred to [62] for more details on this classification method.

Note that the decision regions of trained classifiers can partly overlap. Formally, an ambiguity is given if for an input motor data vector α_t more than two classifiers return +1. To handle this, all N_C classifiers are associated with a weighting factor $w_1, \ldots, w_{N_C}$, being identical to the number of involved data points in each of the N_C trainings. By this, a decision region gets a larger weight when it contains more data points. Each of the K entries of the dictionary $\mathbf{D}$ gets a counter, initialized with zero. It is then iterated over the K atoms: the counter is increased by w_q if $C_q(\alpha_t) = 1$ and has the currently investigated atom in its recommendation. The final decision is then given by choosing those atoms that have the $S_{\max}$ largest weights.

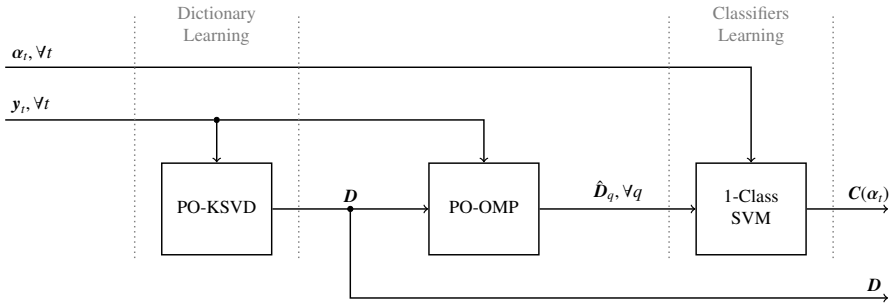

FIGURE 2.6 Audio-motor ego-noise training phase

Illustration of the training phase of the proposed audio-motor integration framework for ego-noise reduction, using training samples $\boldsymbol{y}_t, t = 1, \ldots, T$, and associated motor data $\boldsymbol{\alpha}_t$, $t = 1, \ldots, T$.

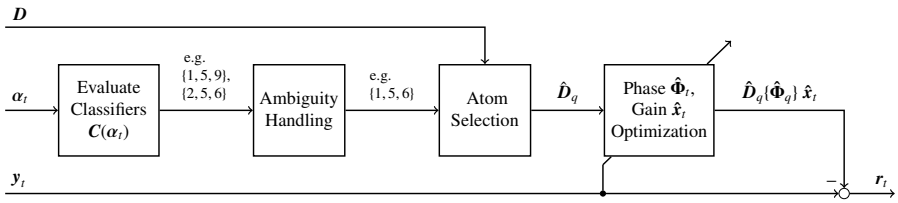

FIGURE 2.7 Audio-motor ego-noise reduction

Illustration of the test-phase of the proposed audio-motor integration framework for ego-noise reduction. Atoms $\hat{\mathbf{D}}_q$ are selected from $\mathbf{D}$ on the basis of motor data $\boldsymbol{\alpha}_t$ only. The incoming audio data sample $\boldsymbol{y}_t$ is a mixture of Ego-noise and a target signal.

2.4.4.2 Method Summary

The proposed proprioceptor-guided multichannel dictionary methodology uses the following steps:

- **Training:** The input consists in spectrogram frame samples $\boldsymbol{y}_t, t = 1, \ldots, T$ containing ego-noise only. Each sample is associated with a motor data vector α_t, $t = 1, \ldots, T$. After $\mathbf{D}$ is learned using PO-KSVD [17] (recall that we do not use motor data for this), PO-OMP is performed with the same samples $\boldsymbol{y}_t$ as input. The selected atoms per sample and their associated motor vector are then processed in the second training step which learns the 1- Class SVMs. This gives N_C classifiers, as defined by (2.7). Each of the classifier is associated to one specific set of atoms from $\mathbf{D}$. Fig. 2.6 gives a schematic overview of the training phase.
- **Testing:** The input consists in a new incoming noisy observation $\boldsymbol{y}_t$ contain-

ing both ego-noise and a target signal to denoise, and a corresponding motor data sample $\boldsymbol{\alpha}_t$. The latter is used to decide immediately on a set of atoms $\hat{\mathbf{D}}_q, q = 1, \ldots, N_C$ using the trained classifiers. The iterative search in the dictionary is unnecessary, so that the proposed algorithm can be expected to be of significantly lower complexity than PO-OMP without motor data. What remains is only the calculation of the gains for all entries in $\hat{\mathbf{D}}_q$, collected in vector form $\hat{\boldsymbol{x}}_t$ and the phase optimization, resulting in the phase matrix $\hat{\boldsymbol{\Phi}}_t$. Determining those unknowns corresponds to the very last step of PO-OMP [17], when all atoms have been selected. Fig. 2.7 gives a schematic overview of the testing phase.

2.4.5 RESULTS

We present an experiment performed with the robot NAO [23] (See Fig. 2.4, left). NAO has four microphones which are all located in the head. Furthermore, the robot has 26 joints, 2 in the head, 12 in the arms, 12 in the legs. We perform exclusively movements of the right arm that involve 6 joints. This gives a feature vector of dimension 12, $\boldsymbol{\alpha} \in \mathbb{R}^{12}$. The sampling frequency for all recordings is $Fs = 16$ kHz, the short-time Fourier transform (STFT) domain uses a Hamming window of length 64 ms and an overlap of 50%. NAO performed its movements in a room with moderate reverberation ($T60 = 200$ ms). Each dictionary used in the following was trained with 30 s of recording. The stationary noise from a cooling fan was removed before the training started. For this, we employed a speech distortion weighted multichannel Wiener filter (MWF) [36, 21]. It needs the power spectral density matrix of pure fan noise as input, which can be easily estimated for constant rotation speed of the fan when the robot is not moving. For testing, 200 utterances from the GRID corpus [10] were recorded with the fan switched off. The loudspeaker was positioned at a 1 m distance of NAO, at a height of 1.5 m. The recorded utterances were added to out-of-training movement noise. These mixtures were then used to evaluate the ego-noise suppression algorithms described above after applying the MWF to suppress the fan noise. The classifiers were trained on 2800 motor data samples in total. To find the best parameter ν and γ, we started a sweep over different settings for both variables. Each setting was cross-checked on a set of data points which was excluded from the training. The overall performance of the ego-noise suppression is measured in terms of Signal-to-Inference-Ratio (SIR in dB) and Signal-to-Distortion-Ratio (SDR in dB), as defined in [71]. While SIR measures the overall noise cancellation, SDR also incorporates information about how much speech is distorted by the suppression algorithm. Additionally, we measure the keyword speech recognition rate (RR), using pocketsphinx [29] in the GRID corpus [10], as defined by the CHiME challenge [70].

We tested different parameter constellations for K and $S_{\max}$. The best results were obtained for a dictionary size of $K = 20$ with sparsity level $S_{\max} = 3$ for this scenario. We parametrized the SVM with a sparsity regularizer $\nu = 2^2$ and a Gaussian kernel width and $\gamma = 2^{-2}$ (see [62] for details on these parameters).

TABLE 2.3 Ego-noise reduction results

Comparison of the proposed audio-motor integrated ego-noise reduction method with two baselines and unprocessed signals, using different metrics.

	SIR [dB]	SDR [dB]	RR [%]
Proposed [60]	**14.71**	**2.64**	**73.0**
PO-OMP [17]	14.46	2.57	71.8
NMF [35]	2.51	0.8	45.2
Unprocessed	-5.48	-8.15	36.1

Table 2.3 compares the results obtained with the proposed audio-motor integration method compared to the results obtained with PO-OMP (no proprioceptor data is used). Both the audio-motor and PO-OMP approach clearly outperform the unprocessed recordings in all metrics used. For comparison, we also give suppression results of one-channel NMF [35]. Although NMF brings an improvement, best results are obtained using PO-OMP and the audio-motor approach. The latter clearly reproduces results of PO-OMP and slightly even outperforms it. This can be explained by the fact that PO-OMP sometimes wrongly estimates atoms due to the presence of speech (recall that PO-OMP uses audio data only). As expected, the needed calculation time in Matlab for the classifiers approach is approximately 30% below that of PO-OMP as the search in the dictionary is unnecessary. Note that the theoretical number of possible atom sequences and therefore classifiers is given by (2.6), i.e., $\binom{20}{3} = 1140$ in our case. Interestingly, only 252 classifiers had to be trained in the given case as only 252 atom sets appeared. Therefore, (2.6) is indeed only an upper bound. Nevertheless, we noted that the computational bottleneck of this approach remains the non-convex estimation of the phase-corrections matrix $\mathbf{\Phi}_t$. This difficult and general problem in audio, referred to as *phase unmixing* is the subject of current research [18].

2.5 CONCLUSION AND PERSPECTIVES

This chapter attempted to bring a unified perspective on audio-motor integration with the key question in mind: *How can audio and motor modalities be combined to enhance robot audition?* After summarizing the challenges and opportunities specific to robot audition, we surveyed the literature on audio-motor integration, from both psychophysics and robotics viewpoints. We then presented two recent examples of learning-based audio-motor integration framework for robotics. In the first one, addressing the fundamental problem of acoustic source localization, the expected derivative of received audio signal log-power in one microphone with respect to motor-state during a motor command is used to derive spatial features. We showed how these features enable 2D sound direction estimation with a single-microphone,

a task impossible to achieve without audio-motor integration. In the second example, addressing the generic problem of ego-noise reduction for acoustic signal enhancement, we presented a general audio-motor integration framework to fuse the information brought by a dictionary-based structured audio model on the one hand, and instantaneous motor data on the other hand. We showed that using motor data reduced the computational complexity while improving performance.

Despite psychophysical evidences pointing out the usefulness of audio-motor integration with humans, only few computational and engineering studies exploit this opportunity. However, with the increasing prominence of robots in our daily lives, and with their fast-developing capability to naturally interact with humans via voice, it is likely that research in computational audio-motor integration will keep gaining momentum. A number of crucial questions will need to be answered along this development:

- How to step out of the classical static-source, static-microphone signal processing framework and, in particular, how to model fast and complex motions?
- What is the optimum auditory representation that can be predicted from motor states? Which motor-space features are best suited for audio prediction?
- How to *close the sensori-motor loop* by deriving long- and short-term optimal actions that ease the enhancement of recorded audio signals?
- Which information may be extracted from dynamic audio inputs for simultaneous localization and mapping, and how to extract them?

We hope that this chapter helps to direct further research interest towards these exciting challenges with great practical relevance on the horizon.

REFERENCES

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. *rmk*-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[2] Sylvain Argentieri, Patrick Danes, and Philippe Souères. A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech & Language*, 34(1):87–112, 2015.

[3] M. Aytekin, C. F. Moss, and J. Z. Simon. A sensorimotor approach to sound localization. *Neural Computation*, 20(3):603–635, 2008.

[4] Hendrik Barfuss and Walter Kellermann. An adaptive microphone array topology for target signal extraction with humanoid robots. In *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*, pages 16–20. IEEE, 2014.

[5] Erik Berglund and Joaquin Sitte. Sound source localisation through active audition. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 653–658. IEEE, 2005.

[6] Mathieu Bernard, Steve NGuyen, Patrick Pirim, Bruno Gas, and Jean-Arcady Meyer. Phonotaxis behavior in the artificial rat psikharpax. In *International Symposium on Robotics and Intelligent Sensors, IRIS2010, Nagoya, Japon*, pages 118–122, 2010.

[7] Mathieu Bernard, Patrick Pirim, Alain de Cheveigné, and Bruno Gas. Sensorimotor learning of sound localization from an auditory evoked behavior. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 91–96. Ieee, 2012.

[8] J Braasch, S Clapp, A Parks, T Pastore, and N Xiang. A binaural model that analyses acoustic spaces and stereophonic reproduction systems by utilizing head rotations. In *The technology of binaural listening*, pages 201–223. Springer, 2013.

[9] Gabriel Bustamante, Patrick Danés, Thomas Forgue, and Ariel Podlubne. Towards information-based feedback control for binaural active localization. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 6325–6329. IEEE, 2016.

[10] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[11] Geoff Davis, Stephane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

[12] Antoine Deleforge, Florence Forbes, and Radu Horaud. Acoustic space learning for sound-source separation and localization on binaural manifolds. *International journal of neural systems*, 25(01):1440003, 2015.

[13] Antoine Deleforge, Florence Forbes, and Radu Horaud. High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911, 2015.

[14] Antoine Deleforge and Radu Horaud. Learning the direction of a sound source using head motions and spectral features. Technical report, INRIA, 2011.

[15] Antoine Deleforge and Radu Horaud. The cocktail party robot: Sound source separation and localisation with an active binaural head. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 431–438. ACM, 2012.

[16] Antoine Deleforge, Radu Horaud, Yoav Y Schechner, and Laurent Girin. Co-localization of audio sources in images using binaural features and locally-linear regression. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(4):718–731, 2015.

[17] Antoine Deleforge and Walter Kellermann. Phase-optimized k-svd for signal extraction from underdetermined multichannel sparse mixtures. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 355–359. IEEE, 2015.

[18] Antoine Deleforge and Yann Traonmilin. Phase unmixing: Multichannel source separation with magnitude constraints. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 161–165. IEEE, 2017.

[19] Christine Evers, Alastair H Moore, and Patrick A Naylor. Acoustic simultaneous localization and mapping (a-slam) of a moving microphone array and its surrounding speakers. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 6–10. IEEE, 2016.

[20] Koutarou Furukawa, Keita Okutani, Kohei Nagira, Takuma Otsuka, Katsutoshi Itoyama, Kazuhiro Nakadai, and Hiroshi G Okuno. Noise correlation matrix estimation for improving sound source localization by multirotor uav. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 3943–3948. IEEE, 2013.

[21] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730, 2017.

[22] Clément Gaultier, Saurabh Kataria, and Antoine Deleforge. Vast: The virtual acoustic space traveler dataset. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 68–79. Springer, 2017.

[23] David Gouaillier, Vincent Hugel, Pierre Blazevic, Chris Kilner, Jérôme Monceaux, Pascal Lafourcade, Brice Marnier, Julien Serre, and Bruno Maisonnier. Mechatronic design of nao humanoid. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 769–774. IEEE, 2009.

[24] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[25] P. M. Hofman and A. J. Van Opstal. Spectro-temporal factors in two-dimensional human sound localization. *JASA*, 103(5):2634–2648, 1998.

[26] Paul M Hofman, Jos GA Van Riswick, and A John Van Opstal. Relearning sound localization with new ears. *Nature neuroscience*, 1(5):417–421, 1998.

[27] Jonas Hornstein, Manuel Lopes, JoschAD:é Santos-Victor, and Francisco Lacerda. Sound localization for humanoid robots-building audio-motor maps based on the hrtf. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 1170–1176. IEEE, 2006.

[28] Jie Huang, Noboru Ohnishi, and Noboru Sugie. Building ears for robots: sound localization and separation. *Artificial Life and Robotics*, 1(4):157–163, 1997.

[29] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alexander I Rudnicky. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

[30] Gökhan Ince, Kazuhiro Nakadai, Tobias Rodemann, Yuji Hasegawa, Hiroshi Tsujino, and Jun-ichi Imura. Ego noise suppression of a robot using template subtraction. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 199–204. IEEE, 2009.

[31] Akinori Ito, Takashi Kanayama, Motoyuki Suzuki, and Shozo Makino. Internal noise suppression for speech recognition by small robots. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[32] Masaharu Kato, Hisashi Uematsu, Makio Kashino, and Tatsuya Hirahara. The effect of head motion on the accuracy of sound localization. *Acoustical science and technology*, 24(5):315–317, 2003.

[33] Laurent Kneip and Claude Baumann. Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis. *The Journal of the Acoustical Society of America*, 124(5):3108–3119, 2008.

[34] Miranda Kreković, Ivan Dokmanić, and Martin Vetterli. Echoslam: Simultaneous localization and mapping with acoustic echoes. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 11–15. Ieee, 2016.

[35] Yifeng Li and Alioune Ngom. Versatile sparse matrix factorization and its applications in high-dimensional biological data analysis. In *IAPR International Conference on Pattern Recognition in Bioinformatics*, pages 91–101. Springer, 2013.

[36] Heinrich W Löllmann, Hendrik Barfuss, Antoine Deleforge, Stefan Meier, and Walter Kellermann. Challenges in acoustic signal enhancement for human-robot communication. In *Speech Communication; 11. ITG Symposium; Proceedings of*, pages 1–4. VDE, 2014.

[37] Heinrich W Löllmann, AlastairH Moore, Patrick A Naylor, Boaz Rafaely, Radu Horaud, Alexandre Mazel, and Walter Kellermann. Microphone array signal processing for robot audition. In *Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017*, pages 51–55. IEEE, 2017.

[38] Ning Ma, Tobias May, and Guy J Brown. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2444–2453, 2017.

[39] Ning Ma, Tobias May, Hagen Wierstorf, and Guy J Brown. A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2699–2703. IEEE, 2015.

[40] Aly Magassouba, Nancy Bertin, and François Chaumette. Sound-based control with two microphones. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 5568–5573. IEEE, 2015.

[41] Tobias May, Ning Ma, and Guy J Brown. Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2679–2683. IEEE, 2015.

[42] John C Middlebrooks and David M Green. Sound localization by human listeners. *Annual review of psychology*, 42(1):135–159, 1991.

[43] Kazuhiro Nakadai, Tino Lourens, Hiroshi G Okuno, and Hiroaki Kitano. Active audition for humanoid. In *AAAI/IAAI*, pages 832–839, 2000.

[44] Kazuhiro Nakadai, Hiroshi G Okuno, and Hiroaki Kitano. Real-time sound source localization and separation for robot audition. In *Seventh International Conference on Spoken Language Processing*, 2002.

[45] Kazuhiro Nakadai, Hiroshi G Okuno, and Hiroaki Kitano. Robot recognizes three simultaneous speech by active audition. In *Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on*, volume 1, pages 398–405. IEEE, 2003.

[46] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. Design and implementation of robot audition system HARK-open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[47] Patrick Naylor and Nikolay D Gaubitch. *Speech dereverberation*. Springer Science & Business Media, 2010.

[48] Quan V Nguyen, Francis Colas, Emmanuel Vincent, and François Charpillet. Long-term robot motion planning for active sound source localization with monte carlo tree search. In *Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017*, pages 61–65. IEEE, 2017.

[49] J Kevin O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–973, 2001.

[50] Makoto Otani, Tatsuya Hirahara, and Shiro Ise. Numerical study on source-distance dependency of head-related transfer functions. *The Journal of the Acoustical Society of America*, 125(5):3253–3261, 2009.

[51] Kazuhiro Otsuka, Shoko Araki, Kentaro Ishizuka, Masakiyo Fujimoto, Martin Heinrich, and Junji Yamato. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 257–264. ACM, 2008.

[52] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010.

[53] Stephen Perrett and William Noble. The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America*, 102(4):2325–2332, 1997.

[54] H. Poincaré. *The foundations of science; Science and hypothesis, the value of science, science and method*. New York: Science Press, 1929. Halsted, G. B. trans. of La valeur de la science, 1905.

[55] Alban Portello, Patrick Danes, and Sylvain Argentieri. Acoustic models and kalman filtering strategies for active binaural sound localization. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 137–142. IEEE, 2011.

[56] Rajkishore Prasad, Hiroshi Saruwatari, and Kiyohiro Shikano. Robots that can hear, understand and talk. *Advanced Robotics*, 18(5):533–564, 2004.

[57] Caleb Rascon and Ivan Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210, 2017.

[58] Jordi Sanchez-Riera, Xavier Alameda-Pineda, Johannes Wienke, Antoine Deleforge, Soraya Arias, Jan Čech, Sebastian Wrede, and Radu Horaud. Online multimodal speaker detection for humanoid robots. In *Humanoid Robots (Humanoids), 2012 12th IEEE-RAS International Conference on*, pages 126–133. IEEE, 2012.

[59] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):971–982, 2013.

[60] Alexander Schmidt, Antoine Deleforge, and Walter Kellermann. Ego-noise reduction using a motor data-guided multichannel dictionary. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 1281–1286. IEEE, 2016.

[61] Alexander Schmidt, Heinrich Loellmann, and Walter Kellermann. A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[62] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[63] Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180. IEEE, 2003.

[64] J. W. Strutt (Lord Rayleigh). On the perception of the direction of sound. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 83(559):61–64, 1909.

[65] Ronen Talmon, Israel Cohen, and Sharon Gannot. Supervised source localization using diffusion kernels. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 245–248. IEEE, 2011.

[66] Willard R Thurlow, John W Mangels, and Philip S Runge. Head movements during sound localization. *The Journal of the Acoustical society of America*, 42(2):489–493, 1967.

[67] Vladimir Tourbabin, Hendrik Barfuss, Boaz Rafaely, and Walter Kellermann. Enhanced robot audition by dynamic acoustic sensing in moving humanoids. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5625–5629. IEEE, 2015.

[68] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[69] Jean-Marc Valin, Shun'ichi Yamamoto, Jean Rouat, François Michaud, Kazuhiro Nakadai, and Hiroshi G Okuno. Robust recognition of simultaneous speech by a mobile robot. *IEEE Transactions on Robotics*, 23(4):742–752, 2007.

[70] Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni. The second chimespeech separation and recognition challenge: An overview of challenge systems and outcomes. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 162–167. IEEE, 2013.

[71] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[72] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

[73] Hans Wallach. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4):339, 1940.

[74] D. Wang and G. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. IEEE Press, 2006.

[75] Lin Wang and Andrea Cavallaro. Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles. In *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*, pages 152–158. IEEE, 2016.

[76] Frederic L Wightman and Doris J Kistler. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105(5):2841–2853, 1999.

[77] B. A. Wright and Y. Zhang. A review of learning with normal and altered sound-localization cues in human adults. *International journal of audiology*, 45(S1):92–98, 2006.

[78] Tian Xiao and Qing Huo Liu. Finite difference computation of head-related transfer function for human hearing. *The Journal of the Acoustical Society of America*, 113(5):2434–2441, 2003.